Possums

David Parsons

March 20, 2017

[Rutgers Data 101 Blog](http://data101.cs.rutgers.edu/?q=blog)

[My Data Exploration Homepage](home.html)

[Rutgers Data 101 Homepage](http://data101.cs.rutgers.edu/)

## Background

My name is David Parsons, I am currently a Junior at Rutgers University majoring in Computer Science. The intent of this project was to continue to expose my self to data exploration/analysis and the R programming language. I've completed several posts like this already that you can check out [here](home.html).

## Insipiration

This data set was chosen by my friend, and was cited as the most interesting of the data sets that they had to choose from. I also have never thought that much about possums, so I figured it might be interesting to learn some more about them.

A did a quick Google search and here are some quick facts that I learned about them before I analyzed the data so that I could have some context. The data specifically involved the Common Brushtail Possum:

* Native to mostly Easter Australia
* Head to Body of 32 - 58 cm
* Tail of 24 - 40 cm
* Weight of 1.2 - 4.5 kg
* Males tend to be slightly bigger than females
* 16 - 18 day gestation period
* Up to 13 year lifespan

You can read more about them [here](https://en.wikipedia.org/wiki/Common_brushtail_possum).

## The Data

Data Collected from <https://vincentarelbundock.github.io/Rdatasets/datasets.html>, a repo of random data sets. This data set had been curated as a part of a research study on the morphological variation of 104 mountain brushtail possums, trapped at seven sites from Southern Victoria to central Queensland.

pos <- read.csv("possum.csv")  
summary(pos)

## X case site Pop sex   
## A1 : 1 Min. : 1.00 Min. :1.000 other:58 f:43   
## A2 : 1 1st Qu.: 26.75 1st Qu.:1.000 Vic :46 m:61   
## A3 : 1 Median : 52.50 Median :3.000   
## A4 : 1 Mean : 52.50 Mean :3.625   
## AD1 : 1 3rd Qu.: 78.25 3rd Qu.:6.000   
## BB13 : 1 Max. :104.00 Max. :7.000   
## (Other):98   
## age hdlngth skullw totlngth   
## Min. :1.000 Min. : 82.50 Min. :50.00 Min. :75.00   
## 1st Qu.:2.250 1st Qu.: 90.67 1st Qu.:54.98 1st Qu.:84.00   
## Median :3.000 Median : 92.80 Median :56.35 Median :88.00   
## Mean :3.833 Mean : 92.60 Mean :56.88 Mean :87.09   
## 3rd Qu.:5.000 3rd Qu.: 94.72 3rd Qu.:58.10 3rd Qu.:90.00   
## Max. :9.000 Max. :103.10 Max. :68.60 Max. :96.50   
## NA's :2   
## taill footlgth earconch eye   
## Min. :32.00 Min. :60.30 Min. :40.30 Min. :12.80   
## 1st Qu.:35.88 1st Qu.:64.60 1st Qu.:44.80 1st Qu.:14.40   
## Median :37.00 Median :68.00 Median :46.80 Median :14.90   
## Mean :37.01 Mean :68.46 Mean :48.13 Mean :15.05   
## 3rd Qu.:38.00 3rd Qu.:72.50 3rd Qu.:52.00 3rd Qu.:15.72   
## Max. :43.00 Max. :77.90 Max. :56.20 Max. :17.80   
## NA's :1   
## chest belly   
## Min. :22.0 Min. :25.00   
## 1st Qu.:25.5 1st Qu.:31.00   
## Median :27.0 Median :32.50   
## Mean :27.0 Mean :32.59   
## 3rd Qu.:28.0 3rd Qu.:34.12   
## Max. :32.0 Max. :40.00   
##

cols <- colnames(pos)

Each Data Point (Possum) had the following properties:

X - unique id (possum name...sorta)  
case - observation number  
site - one of seven locations where possums were trapped  
Pop - a factor which classifies the sites as Vic Victoria, other New South Wales or Queensland  
sex - a factor with levels f female, m male  
age - age  
hdlngth - head length  
skullw - skull width  
totlngth - total length  
taill - tail length  
footlgth - foot length  
earconch - ear conch length  
eye - distance from medial canthus to lateral canthus of right eye  
chest - chest girth (in cm)  
belly - belly girth (in cm)

I know nothing about possums besides the normal stereotypical Hollywood persona, so I'm interested to see what I learn from looking at these animals in a more complex manner.

## Helper Funtions

scatPlot <- function(x, y, frm){  
  
 title = paste(x,y, sep = " vs ")  
 plot(frm[,x], frm[,y], col = rainbow(8), main = title, las = 2, xlab = x, ylab = y)  
}

## Investigation

The following is the step by step thought process I followed in exploring this data set and arriving at the conclusion I did.

### Scatter Plots

I started by plotting every column against every other column in an effort to visual the data and make it easier to confirm any biases or recognize any unique patterns. I don't recommend running this code on your computer, especially if it is under powered.

nos <- c("X", "case", "Pop", "site")  
for(x in cols){  
 if(x %in% nos){  
 next  
 }  
 for(y in cols){  
 if(y %in% nos){  
 next  
 }  
 if(y != x){  
 scatPlot(x, y, pos)  
 }  
 }  
}

From this I produced the following observation notes:

commons <- c("belly", "chest", "footlgth", "totlgth", "skullw", "hdlngth")  
Linear Trend for the following  
belly v common  
chest v common  
footlgth v common  
totlngth v common  
hdlngth v common  
  
Ear Conch Groups  
  
Age pyramid  
  
Females Bigger?  
 Longer Really  
 And older

## Trend Investitgation

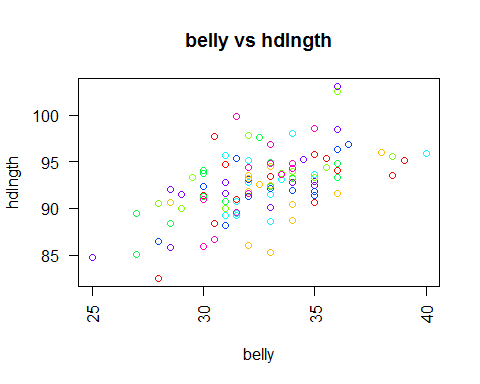
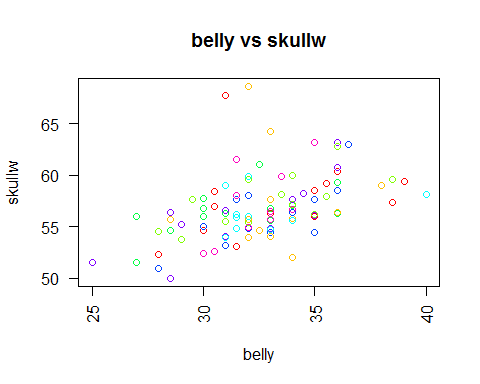
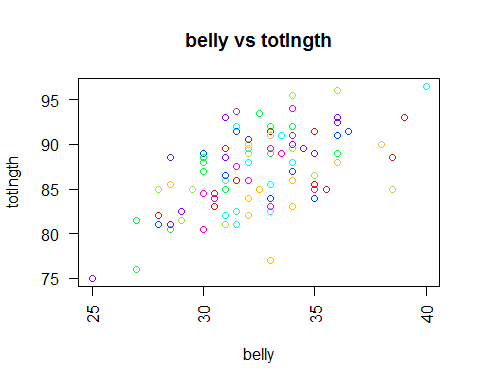
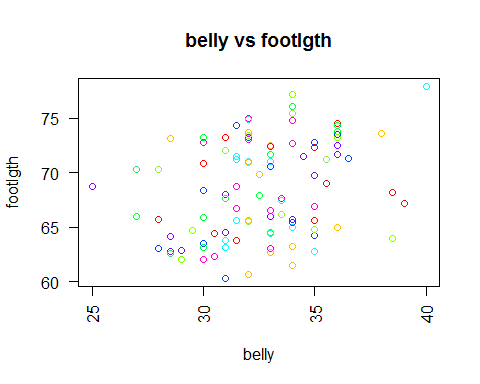
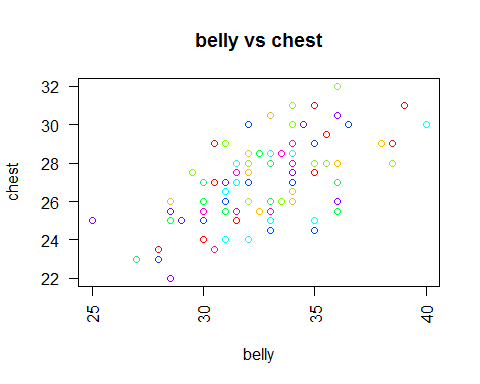
No that I have made some observations, I think its best to focus on each one, analyze it alone and determine if it's believable or not.

### Trend 1 - The Linear Commons

The commons were coined during my investigation, because any common plotted against any other common yielded similar linear distributions.

The commons consist of belly, chest, footlgth, totlgth, skullw, and hdlngth.

commons <- c("belly", "chest", "footlgth", "totlngth", "skullw", "hdlngth")  
  
x = "belly"  
  
for(y in commons){  
 if(x != y){  
 scatPlot(x, y, pos)  
 }  
}



I think that this is very believable and doesn't really warrant anymore investigation.

Bigger possums are, well, bigger. As a possum increases in size of one common, they increase in size in all commons. This makes sense, as most of these commons are just body attributes. So the bigger the possum the bigger the commons.

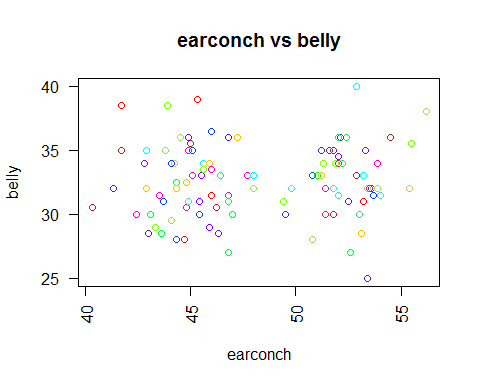
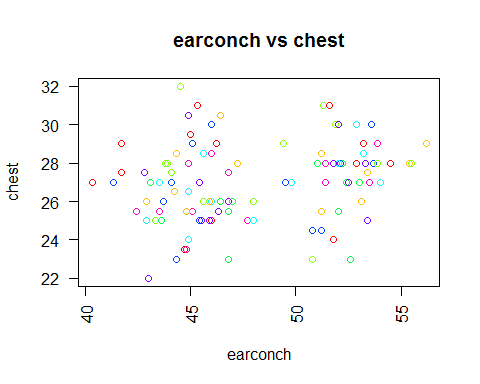
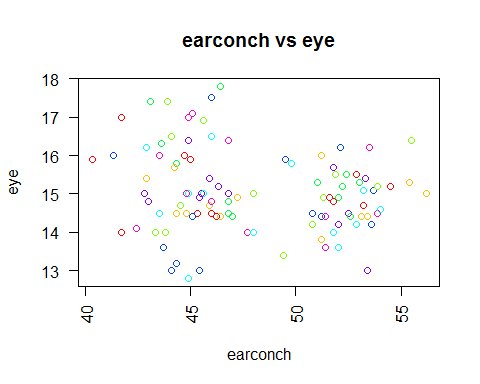
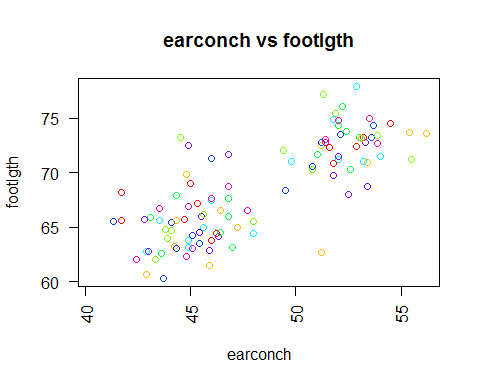
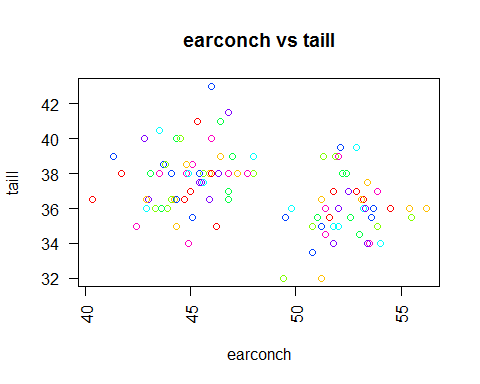
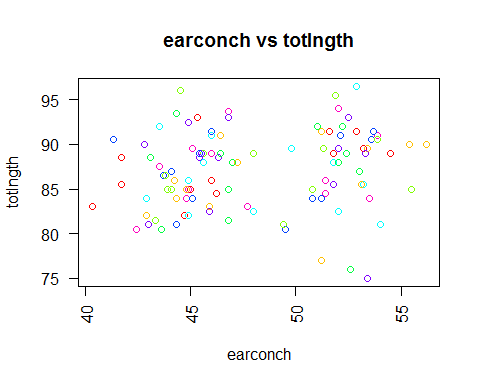
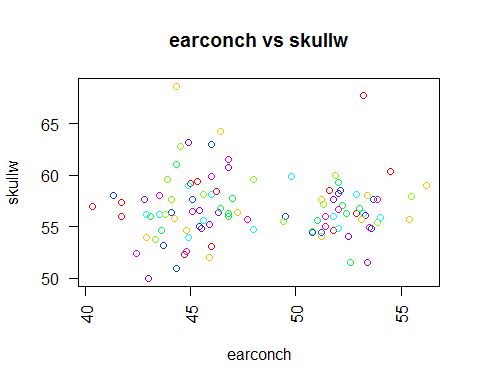
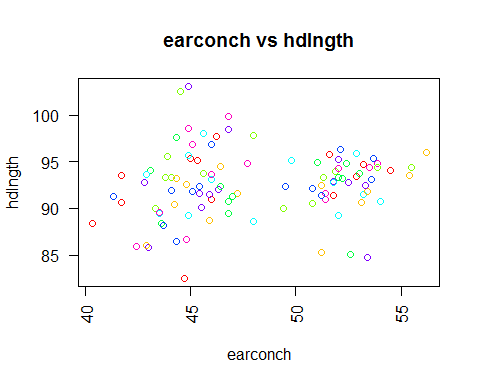
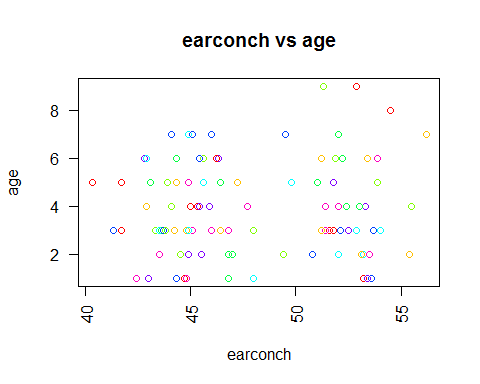
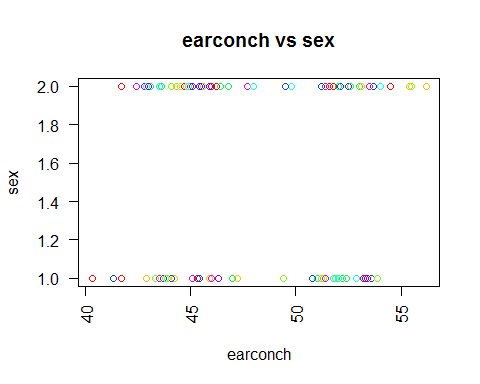
I won't do any p-testing here, since well this is common sense.

### Trend 2 - Short vs Long Ear Conch

The commons were coined during my investigation, because any common plotted against any other common yielded similar linear distributions.

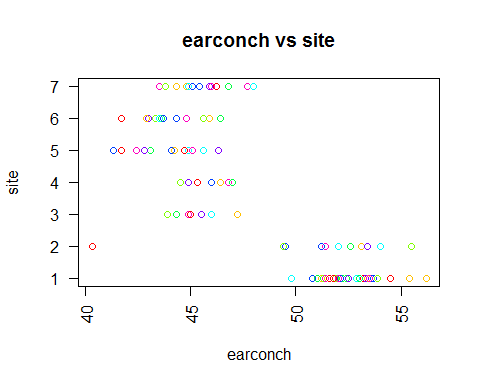
The commons consist of belly, chest, footlgth, totlgth, skullw, and hdlngth.

x = "earconch"  
  
for(y in cols){  
 if(x != y && y != "X" && y != "case" && y != "site" && y != "Pop"){  
 scatPlot(x, y, pos)  
 }  
}



There is a clear separation in earconch sizes in all other categories. Begging the question as to what is causing these separations.

scatPlot("earconch", "site", pos)



Here it is easy to see that there sites 1 & 2 were locations where possums with larger ear conch lengths reside. This would explain the separation in each category, and the normal distribution otherwise.

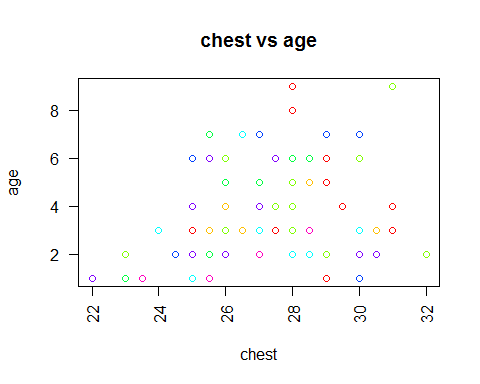
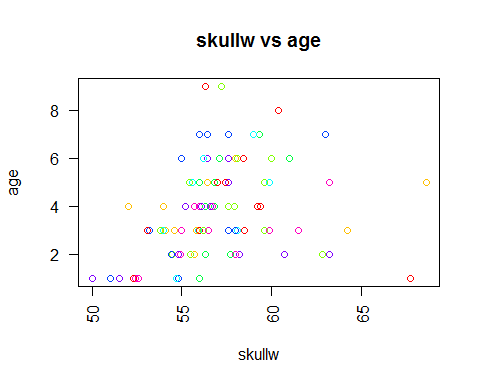
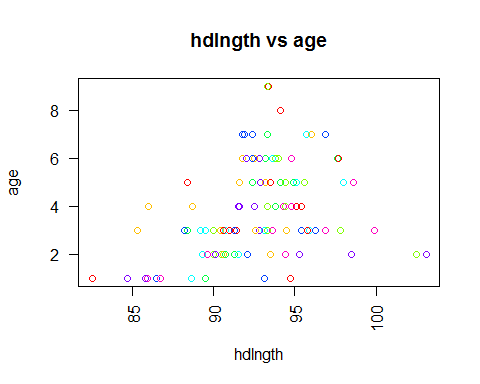
Again, I'm not an expert on possums. But perhaps there are two different variants of this species of possum, a long and short ear.

I don't think this warrants a p-test either, given the type of conclusion.

### Trend 3 - The Age Pyramid

You may have noticed that age was left off the commons. Not only did it not follow the common pattern. But it also had some interesting behavior of its own.

x = "age"  
  
for(y in c("hdlngth", "skullw", "chest")){  
 scatPlot(y, x, pos)  
}



Head Length, Skull Width, and Chest Size show this pyramid I noticed in preliminary plotting.

As possums age, so does their value in head, skull, and chest, which makes sense since possums grow as they get older. But this stops at a point, and you begin to see these value continue to increase but the age starts to decrease. Some of these younger possums are bigger than the older ones.

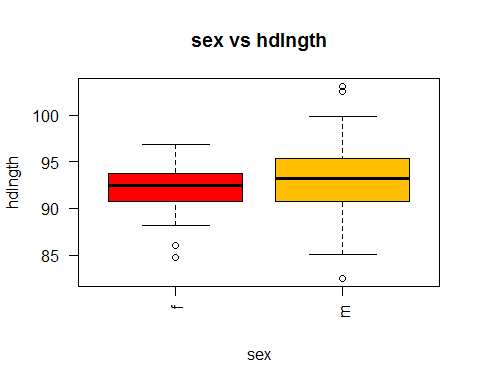
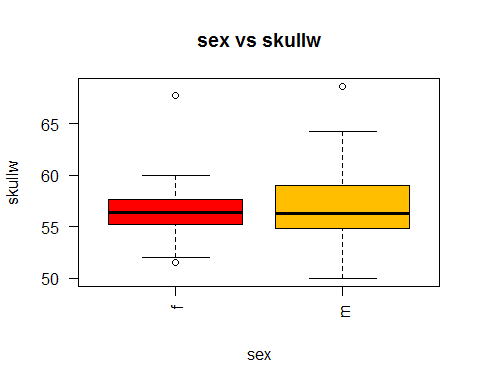
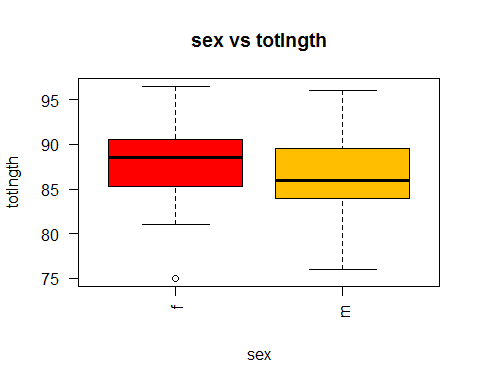
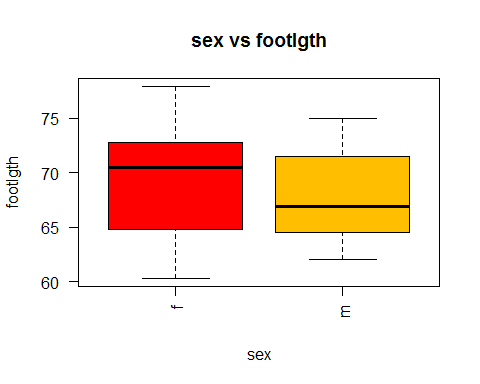
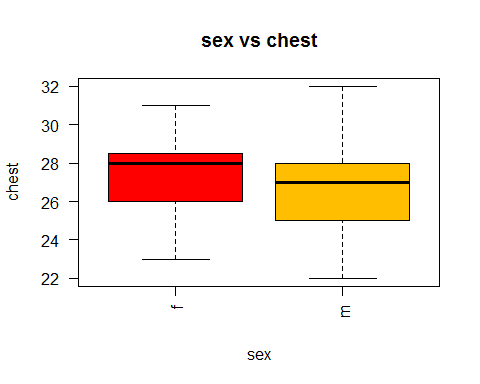
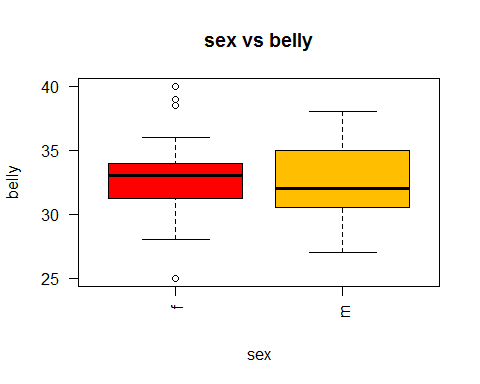
That's weird. Are the possums shrinking? Perhaps there is some body deterioration in older possums. But I find this hard to believe so given that this trend doesn't occur for all body measurements, I am willing to accept that it's random.

### Trend 4 - Females Defying Gender Roles

If you looked at introduction to this post, or read the linked Wikipedia page on common bushtail possum, or for some other reason are up to date on your possum facts, then you know that male possums tend to be larger in size than their female counterparts.

Well the following trend calls this into question.

x = "sex"  
  
for(y in commons){  
 if(x != y && y != "X" && y != "case" && y != "site" && y != "Pop"){  
 scatPlot(x, y, pos)  
 }  
}



In this data set, it appears that females possums are actually somewhat larger than males on average.

Given that this observation directly contradicts the research on the animal, I've decide to run the mean of each measurement through some permutation tests.

pos[, 16] <- ""  
colnames(pos)[16] <- "isFemale"  
pos$isFemale <- (pos$sex == "f")  
  
for(y in commons){  
 if(y != "X" && y != "case" && y != "site" && y != "Pop" && y != "footlgth"){  
 print(PermutationTestSecond::Permutation(pos, "isFemale", y, 50000, FALSE, TRUE))  
 }  
}

## [1] 0.17612  
## [1] 0.07626  
## [1] 0.05166  
## [1] 0.21426  
## [1] 0.14006

And none of these p values really me allow me to accept an 95% confidence interval. So I'm going to go ahead and say that it is likely that this data just happens to reflect that female possums are larger than male possums by chance.

There could be some correlation between the average age of male possums and female possums that were sampled, that could explain this phenomena.

## Conclusion

This analysis didn't provide me with any eye opening information about Common Bushtail Possums in general. And given that this data came from a select sampled population of the species in a specific location, that isn't very surprising.

But this data shows a few interesting things about the species relative to the area where this data was sampled.

There appears to be predominately two types of possums in the area, categorized by their ear conch length. Some further investigation may show the reason why this might be the case.

It also appears that female possums may defy their typical gender roles in this area, signaling that either the culture in the area is quite progressive, or that the data is biased.

And the effects that time has on the physical statures of possums might also be an interesting area of inquiry as it appears that possums start to shrink as they get older.

## Notes and Comments

While I didn't start this project thinking that I would conduct it on possums, and have enjoyed working with this data. In my only other previous work with a real world data set, I worked on relationships that were still relatively theoretical, but with this data set I found myself thinking about possums a lot and the habits that might produce the results I was coming across. This type of imagery grasped my interest as I worked, but may have influenced my analysis.

This was also the first time I used p-testing on real data. I really appreciate this method of challenging hypothesis that are presented, as it only opened more doors of exploration, rather than shutting them down by just assuming that I was correct.

I hope that in future explorations I can do my analysis and testing more efficiently and potentially generate some strong real world arguments, and who knows, maybe I'll do it on possums again.

[Rutgers Data 101 Blog](http://data101.cs.rutgers.edu/?q=blog)

[My Data Exploration Homepage](home.html)

[Rutgers Data 101 Homepage](http://data101.cs.rutgers.edu/)